RNA structure prediction: From 2D to 3D

Bernhard C. Thiel\(^1\) Christoph Flamm\(^1\)
Ivo L. Hofacker\(^{1,2,}\)*

July 19, 2017

\(^1\)Department of Theoretical Chemistry, Faculty of Chemistry, University of Vienna, Austria
\(^2\)Research Group BCB, Faculty of Computer Science, University of Vienna, Austria

*To whom correspondence should be addressed.

Abstract
We summarize different levels of RNA structure prediction, from classical 2D structure to extended secondary structure and motif based research towards 3D structure prediction of RNA. We outline the importance of classical secondary structure during all those levels of structure prediction.

1 Introduction
The secondary structure model of RNA is at the center of most computational work related to RNA. In this review we will sketch how recent research has taken advantage of the increasing amount of experimentally solved RNA 3D structures to move beyond the secondary structure view, but how it cannot move away from it.

RNA secondary structure is defined by canonical AU, GC base-pairs and GU wobble-pairs, which form between nucleotides typically within one RNA strand to create anti-parallel A-type helices. Pseudoknot-free secondary structures can be encoded as strings, e.g. in the Vienna dot-bracket notation, where nested pairs of left and right parentheses signify nested base pairs, while dots signify unpaired nucleotides (See Figure 1, box “Secondary Structure”).

RNA secondary structure can be efficiently predicted from the primary sequence using dynamic programming approaches. The easiest such algorithm maximizes the number of base-pairs for a given sequence \([1]\), while state-of-the-art prediction tools \([2]\) use an energy model with different energy contributions for different loop types. This model takes the stacking of neighboring base pairs into account and is often referred to as nearest neighbor model. The energy parameters of the nearest neighbor model were derived from RNA melting experiments collected by Turner et al \([3]\). These secondary structure prediction algorithms can be used not only to predict the structure with the lowest free energy, but also to generate the Boltzmann ensemble of suboptimal structures \([2, 4, 5]\). By including data from chemical probing experiments such as SHAPE, the reliability of the generated secondary structures can be further increased. For an overview over state of the art techniques in secondary structure prediction and their integration with chemical probing data, see the recent review by \([6]\)
Lorenz et al. [6]. Pseudoknots increase the computational complexity, but simple types can be predicted with dedicated software [7].

The great success of the secondary structure description of RNA is grounded in the hierarchical nature of RNA folding [8]. Intra-helical interactions have a significantly greater energy contribution than tertiary interactions. Thus, energy barriers and in most cases energy differences between different secondary structures are significantly higher than energy differences and barriers between different 3D structures that correspond to the same secondary structure.

Secondary structure is useful in many approaches beyond simple structure prediction: Sequence design [9, 10] aims at finding sequences that fold into one or more predefined (meta-)stable structures. Folding kinetics [11] and co-transcriptional folding [12, 13] look at the dynamic nature of RNA secondary structure formation. Functional RNA secondary structures are expected to be evolutionarily conserved. Hence they can be detected in genomic data via compensatory mutations and structure stability [14] or co-variance models (RNA families) [15, 16].

While some biological mechanisms can be understood on the basis of secondary structure alone, others require more detailed structure knowledge. In this contribution we will therefore focus on approaches that go beyond classical secondary structure.

On the other end of the spectrum lie methods that model RNA structures in full atomistic resolution. In particular, Molecular Dynamics simulations can be readily applied to RNA, with the AMBER force-field being most popular in the RNA community. However, despite many RNA related corrections to the AMBER force field, the current version still sometimes fails to identify the correct native state among all possible states [17, 18]. Despite these challenges, reduced model [19] and explicit solvent MD simulations [20] of the full ribosome have been successfully performed, using experimental crystal structures as a starting point and, in the case of the reduced model, for the construction of a structure-based (Gō-like) potential. The reduced model simulation covered biologically relevant time-scales, while the explicit solvent MD simulation covered over a microsecond, enough to separate rapid local fluctuations from large-scale collective movements. The latter simulation took several months on over 1000 computer cores.

In contrast to molecular dynamics from a given starting conformation, true de-novo simulations are possible only for very small molecules, due to the huge space of possible conformations and rugged energy landscapes, even when combined with enhanced sampling techniques [21]. Recent applications of classical force fields to RNA folding range from the folding of G-quadruplexes [22] and RNA-ion interactions [23] to the folding of tetra-loop hairpins [17].

While all-atom approaches are limited to small molecules or fragments, somewhat larger RNAs can be treated by methods working with a coarse grained RNA representation, which we will review in sections [8] and [4]. In both cases, prior knowledge of the secondary structure is not required. Rather, correct secondary structure should emerge as a by-product of tertiary structure prediction. Occasionally, known secondary structures are used as constraints to reduce the search space. However, we will also discuss tertiary structure methods that explicitly build on secondary structure knowledge and use it to enable a more aggressive coarse-graining.

For a discussion on different ways of coarse-graining in RNA 3D structure prediction
Figure 1: RNA structure prediction programs can be classified into hierarchical folding approaches and all-in-one programs.

prediction and the distinction between theory based and knowledge based potentials, see the recent review by Dawson et al. [24].

2 Extended secondary structure

While prediction of full tertiary RNA structures remains an extremely difficult task, several promising avenues have emerged to go beyond classical secondary structure without tackling actual 3D structure, e.g. by extending the notion of secondary structure to include non-canonical interactions. In addition to the Watson-Crick and GU wobble pairs, RNA nucleotides can form a wide variety of interactions, both pairwise and between more than two nucleotides.

A highly useful classification of these non-canonical interactions was introduced by Leontis and Westhof [25], by noting that almost all interactions happen on one of 3 “edges” (the Watson-Crick, Hoogsteen, and sugar edge). Together with the relative orientation of the glycosidic bond (cis or trans) this results in 12 basepairing types, each of which has at least two hydrogen bonds and was observed for at least some combinations of bases [26]. Additionally, weaker interactions involving a single hydrogen bond, base triplets or multiplets and G-quadruplexes have been observed in RNA molecules.

An RNA structure including some or all of the above-mentioned non-canonical interactions, but not the full 3D information, will be called an extended secondary structure. Decomposition and motif search are the two main approaches for the prediction of extended secondary structures.
Extended secondary structures play an important role in many natural RNAs, including ribosomal RNAs\cite{27}, where non-canonical basepairs contribute to the movement of the ribosome\cite{28}.

2.1 Loop decomposition with non-canonical basepairs

Classical secondary structure prediction is based upon the unique decomposition of the RNA structure into loops delimited only by canonical and G-U basepairs. The program MC-Fold\cite{29} generalizes the classical decomposition by considering all types of non-canonical basepairs as loop delimiters. The resulting loops, termed NCMs (Nucleotide cyclic motifs) in MC-Fold, are defined via a minimal cycle basis of the RNA structure graph\cite{30}.

The programs MC-Fold\cite{29} and the faster dynamic programming implementation MC-Fold-dp\cite{31} find optimal and sub-optimal combinations of NCMs for a given sequence according to a statistical energy function. This energy function is composed of probability terms for finding individual and combinations of two NCMs.

RNAwolf\cite{31} further generalizes the MC-Fold approach by allowing each nucleotide to form two interactions, thus supporting base-triplets which are observed quite frequently in nature. RNAwolf, however, uses a simplified Nussinov-like energy model, limiting its accuracy. Parametrization of a full featured energy model remains the big challenge for extended secondary structure prediction, and it is unclear if enough data is available to estimate the large number of parameters required. Moreover, while the nearest neighbor approximation has been shown to be quite accurate for canonical secondary structures, the distortion of the double helix through non-canonical pairs could well exert an influence beyond its direct neighbors.

2.2 Motif based

An alternative to the loop decomposition above is to treat substructures containing several non-canonical pairs as a single unit, called a tertiary structure motif. Such an RNA motif is characterized by well-defined (base-pairing) interactions\cite{32,33}, a well-defined geometry\cite{34,35,36,37}, or both. The prime example of such a recurrent motif would be the kink turn motif, a sharp turn between two adjacent helices introduced by a number of bases forming a dense network of non-canonical interactions. This and three other common motifs can be detected in single sequences or multiple sequence alignments using RMDetect\cite{38}.

A well-maintained collection of such motifs can be found in the Motif Atlas\cite{39}. An accompanying search tool, JAR3D\cite{40}, can be used to predict the presence of any motif from the Atlas in interior or hairpin loops of a secondary structure.

Motif based approaches circumvent the problem of decomposing complex interaction networks, such as the kink turn, and even allow for crossing interaction within these motifs. On the downside, they cannot predict novel motifs and are limited to the set of motifs that we observe in known tertiary structures. One application of motifs is homology modeling, where non-homologous regions can be filled in using motifs\cite{41}.
Prediction of G-quadruplexes can also be seen as a motif based extension of secondary structure. G-quadruplex prediction has recently been integrated in the folding algorithms of the ViennaRNA package [42]. This integration presents a significant advantage over pure sequence searches, as it properly treats the competition between formation of G-quads and normal secondary structure.

Extended secondary structures are of great interest because functionally important interaction with proteins and other factors, typically take place in regions of irregular structure, rather than perfect helices. Moreover, extended secondary structures provide a much better starting place for modeling tertiary structures. The kink turn again serves as a perfect example: The sharp bend introduced by this motif will strongly affect the overall shape of the molecule, but would be very hard to predict by any approach that is not aware of the motif.

3 Hierarchical folding enables aggressive coarse-graining

While secondary structure can be used in most 3D structure prediction programs to constrain the search space, for some programs the secondary structure is at the core of the abstraction, on which the coarse grained structure representation is based.

Vfold3D [43, 44] creates a coarse grained 3D scaffold from the RNA secondary structure and sequence using a template based approach. For each loop region, the best matching template is selected from a template library constructed from PDB structures, where the match quality is based on sequence similarity. Helices are modeled as ideal A-type helices. From this scaffold, a full atom model is constructed which is then relaxed using an AMBER all-atom force field.

More than one fragment can be tried for each loop in an exhaustive way (5 fragments were tried in the RNA puzzles entries - see below). However, Vfold3D does not contain any energy function to score these coarse-grained scaffolds, nor does it contain a sampling protocol to sample combinations of such fragments. Since it is unlikely that AMBER energy minimization will overcome larger energy barriers, the quality of the prediction depends on the correct choice of fragments. Thus Vfold works best when the structure of loops strongly depends on the sequence (see section 2.2 about motifs) or Vfold contains close homologs of the target RNA in its knowledge-base, while it has limitations whenever similar loop sequences can adopt multiple loop conformations.

RNAComposer [45] uses a machine translation system that translates RNA
sequence and secondary structure into 3D structure. Similar to Vfold it selects the best matching fragment for every loop and helix from a database of fragments. RNAComposer provides fallback mechanisms for cases where no fragment is found as well as two final refinement steps using CYANA (for refinement in torsion angle space) and the CHARMM force field (for refinement in cartesian space). Like Vfold3D, it allows for the creation of random suboptimal structures but does not provide sampling methods or a fast to evaluate energy function for a systematic exploration of the conformational space.

ERNWIN [46] and RAGTOP [47] use the user-supplied secondary structure to guide their aggressive coarse-graining of the RNA into helices (stems) and connecting loops. In contrast to RNAComposer and VFold3D, these tools explore the conformational space on the level of loops and helices.

RAGTOP [47], first employs machine learning to predict the topology of multiloops (junctions) [48]. Next interior loop angles are sampled with a Monte Carlo/ Simulated Annealing algorithm using a knowledge-based potential. Finally an all-atom representation is recovered from the sampled conformations [49]. With RMSD values from 2.38 to 14.56 (for structures of 25 to 158 nucleotides) the reported prediction accuracy of RAGTOP is slightly better than previous tools which use less aggressive coarse-graining.

ERNWIN [46] uses a similar coarse-graining based on secondary structure. Helices are described by 10 parameters: The position of the helix’s start and end (or equivalently the start and the direction) and 4 parameter for the position of the minor groove along the length of the helix. This model assumes a regular helix where the position of the minor groove with respect to the helices axis changes by a fixed angle with each subsequent nucleotide. The helix axis and the vector pointing to the minor groove form a local coordinate system for residues. Using average atom position data in this local coordinate system gathered from a non-redundant list of PDB files, so-called “virtual atom positions” of all the helix atoms can be quickly calculated just from the 10 parameters defining the helix. Another major difference between RAGTOP and ERNWIN is the fact that the later can sample different multiloop-conformations. This means that the prediction quality of ERNWIN is independent of the correctness of the junction topology prediction, but comes at the cost of a higher number of rejections during Monte Carlo sampling. Finally, sampling and energy evaluation are significantly different between RAGTOP and ERNWIN. In ERNWIN, local loop and helix conformations are sampled directly from a fragment library, whereas RAGTOP samples the angle from continuous space and therefore needs local energy terms to evaluate the likeliness of a local angle. As for the contribution of global features to the energy function, both programs have a term for the radius of gyration. ERNWIN in addition contains a term for loop-loop interactions and a sophisticated term for interactions of single stranded Adenines with the minor groove of a helix (A-Minor motif [50]). ERNWIN also has direct support for motif search via JAR3D (see above).

While the idea of using helices without internal degrees of freedom is common to RAGTOP and ERNWIN, it is the details that matter: The sampling strategy that explores the conformational space and the energy function that detects native-like conformations require a lot of fine-tuning and still have room for improvement.

MC-Sym [29] uses a different approach towards tertiary structure prediction. It is built on top of MC-Fold which can predict extended secondary structures.
It uses a library of 3D fragments for each nucleotide cyclic motif (NCM) predicted by MC-Fold to assemble 3D structures. In a Las Vegas algorithm, 3D structures are sampled for 12 hours by aligning adjacent NCM-3D fragments to form structures. The final result is an ensemble of 3D structures.

4 Predicting secondary and tertiary structure together (“all in one”)

While the tools in the previous section built their model on top of a given secondary structure, the following programs add secondary structure constraints into their model via a force field.

The program NAST \cite{51} represents each nucleotide by a single point, which means that it does not hold any information about the orientation of the base with respect to the backbone. The program requires an input secondary structure, which is used to create energy potentials on the lengths, angles and dihedrals between residues. These potentials direct the sampling of the RNA towards the desired secondary structure. We argue that the requirement for a secondary structure is inherent to the coarse grained model used, because formation of new basepairs would depend on the orientation of the base, which is not part of the model. In NAST, sampling via a Molecular Dynamics approach starting from the extended chain is followed by filtering according to tertiary structure constraints (e.g. from SAXS or SHAPE experiments or from phylogenetic observations \cite{52}) and clustering using an k-means algorithm based on a simplified pairwise GDT-TS \cite{53} distance. For one RNA molecule the effect of errors in the secondary structure input was investigated and the authors concluded that up to 35% wrong base-pairs did not significantly reduce the RMSD of their prediction.

The most aggressive coarse-graining possible that allows for de-novo prediction of base-pairs uses one rigid body per nucleotide (in contrast to one point), as implemented in oxRNA \cite{54, 55} (see below). In a similar fashion, FARNA uses rigid fragments of 1 to 3 nucleotides to assemble the final RNA structure. Next to energy terms for clashes and the radius of gyration, FARNA uses a base-pairing statistical potential that can be seen as a heatmap around the center of an ideal base. If two bases are placed in a relative orientation that is frequently found in base-pairs in solved RNA structures, these bases receive a favorable energy contribution. Structures generated by FARNA can be refined using an all-atom force field, giving rise to the combined method FARFAR \cite{56, 57}. In theory such an all-atom refinement would be useful for all coarse grained approaches (but see introduction for the limits of current force-fields), however its integration in one framework (namely the Rosetta framework) is a great technical advantage of FARNA/FARFAR.

In contrast to FARNA, oxRNA \cite{54, 55} samples structures from a continuous space. It can be used for Molecular Dynamics Simulations and for Monte Carlo calculations. The energy function of oxRNA is parametrized to favor RNA A-type helices by using potentials for sequence dependent hydrogen bonding, stacking and cross-stacking in helices in addition to the backbone potential. Since all nucleotides are equally sized, mismatches in helices do not disturb the overall helix geometry in oxRNA and thus the stability of such helices is over-estimated.
despite the lack of the energy contribution from the hydrogen bonding term. The quality of the secondary structure predicted by oxRNA is comparable with secondary structure prediction tools using the nearest neighbor model for small RNAs, as shown by melting temperature calculations done with oxRNA. Since oxRNA does not include any non-canonical or long-range interactions (except excluded volume effects), it is no surprise that the use of a 3-dimensional model does not bring any significant advantage over a 2-dimensional model with respect to the prediction of secondary structures. The oxRNA paper studied mechanical properties like force-extension and overstretching properties, persistence length and modeled hairpin-unzipping. These applications, while leading to results that are only in the same order of magnitude as the experimental values, could not be done with most of the other RNA structure prediction models.

In contrast to the continuous energy model used in oxRNA, simRNA \cite{58,59} and iFoldRNA \cite{60} use discrete, grid-based statistical potentials. While iFoldRNA uses Discrete Molecular Dynamics \cite{61,62}, simRNA uses Monte Carlo simulations. iFoldRNA \cite{60} uses 3 beads per nucleotide: one for the sugar, one for the phosphate and only one for the base (see Figure 3). The position of the base relative to the sugar is used to determine the direction of hydrogen bonding, but tilting of the base’s plane cannot be modeled with this coarse-graining. Non-canonical interactions are implicitly modeled by the use of a general hydrophobic attraction between all kinds of bases. iFoldRNA’s energy function only contains local terms (for bond length, angle and dihedral angle, base pairing, phosphate-phosphate repulsion, hydrophobic interactions and base stacking). Furthermore an additional energy term for the loop entropy is used to compensate for the bias in loop entropy introduced by the coarse-graining. According to the data reported in the paper’s supplement \cite{60}, iFoldRNA predicts pseudoknot-free secondary structures slightly better than MFold (average Q-value of 0.953 vs 0.948) and can additionally predict pseudoknots. However, we note that the reported benchmark of secondary structure prediction only includes the Q-value (true positives) and does not take false positives (additional basepairs) into account.

SimRNA \cite{58,59} models the base with 3 points and is thus able to capture the full 3-dimensional orientation of the base’s plane. The statistical energy terms for base-base, base-backbone and backbone-backbone interactions can be visualized as heatmaps similar to the ones used in the original FARNA publication \cite{63}. The authors report that their model performs better than other 3D structure prediction programs for short RNA sequences, but needs explicit secondary structure (and potentially long-range) restraints to model longer RNA molecules correctly.

Among the last 3 discussed models, simRNA is the only one which fully incorporates non-canonical base-pairs in its energy model. Since secondary structure constraints are modeled as distance constraints, (extended) secondary structures could in principle be used as input, but the type of the base-pair in the prediction might be different from the type in the input structure.

5 RNA puzzles

As of now, 3 rounds of RNA puzzles, a CASP-like blind experiment in RNA 3D structure prediction, have been completed \cite{64,65,66}. In these competitions,
participating experimental groups provide solved but unpublished tertiary structures. Computational groups then have a few weeks time to perform tertiary structure prediction given the sequence and - in some cases - chemical probing information.

The submitted models are scored using different measures: The root mean square deviation (RMSD) is a common measure for comparison of macromolecules which can be quickly calculated [67], but has received some critiques [68]. The Interaction Network Fidelity (INF) can be calculated for stacking and (non-)canonical hydrogen bonding interactions. While the RMSD is sensitive to errors in flexible loop regions, the INF is sensitive to errors in structured regions. Additionally, RMSD and INF are combined to the Deformation Index (DI). From round II of RNA puzzles onwards, the mean of circular quantities (MCQ) [69] which is based on angular coordinates, was used as additional measure. In contrast to the RMSD, errors in the relative orientation of two relatively large RNA domains do not have a higher impact on the MCQ score than errors in the orientation of small stems or hairpins.

In the first round of RNA puzzles, one riboswitch domain, a stem-loop structure with two interior loops and a square formed by 4 helices and interior loops in the corners were modeled. Of these challenges, only the riboswitch domain contained true higher order 3D folds. Interestingly, a fragment based approach (Vfold) combined with relaxation in an all-atom AMBER force-field yielded the result with the lowest RMSD for this puzzle. This probably means that fragments from homologous RNA molecules were selected based on the sequence similarity score. The second best result of this puzzle used a coarse grained approach with secondary structure constraints, followed by full-atom refinement in a Discrete Molecular Dynamics framework (iFoldRNA). One lesson learned from this RNA puzzle is certainly that the huge size of the sampling space requires some sort of coarse-grained initial step which uses the secondary structure, followed by an all-atom refinement.

In round II [65] of the RNA puzzles contest, SHAPE data was provided to

![Figure 3: Different coarse grained representations of nucleotides.](image)

**NAST** uses a single point located at the C3' atom. **oxRNA** uses a single rigid body with multiple interaction sites defined with respect to the RNA’s center of mass (the positions in the figure are rough estimates of this interaction sites) **iFoldRNA** uses 3 beads per residue, located at the center of the phosphate, the sugar and the base respectively. **simRNA** uses 5 beads per nucleotide, 3 of which define the plane of the base.
all participants by one group. Since this round’s target structures were longer RNA molecules with more complex 3D folds (a ribozyme, a riboswitch and a T-box-tRNA complex), these chemical probing data were crucial for 3D structure prediction. Despite the use of experimental secondary structures, no perfect predictions were obtained, which shows the open challenges in RNA 3D structure prediction.

In round III it became apparent how homology modeling can achieve good results when a homolog with solved structure exists, and how targets without homology to solved structures still pose huge challenges to modelers. It is also notable that many groups who submitted several models could not correctly rank the best prediction as their top choice.

6 Summary

While RNA secondary structure prediction is well established and widely used, tertiary structure prediction has long seemed out of reach. Recently, however, two directions have emerged that promise RNA structure models that go beyond secondary structure. Inclusion of structure motifs and non-canonical base pairs could yield extended secondary structures that are more detailed and will perhaps even improve prediction accuracy. Coarse grained models of tertiary structures may overcome the sampling problems that limit all-atom methods.

A successful RNA 3D structure prediction pipeline, as illustrated in figure 4, will need several ingredients: It should start from a reliable secondary structure, preferably including tertiary motifs and supported by experimental evidence, such as probing data. Exploration of the conformation space will be best done using coarse grained models that provide efficient sampling. Empirical scoring functions need to be able to identify coarse-grained conformation(s) close to the native state. Additional experimental restraints can be incorporated at this step, as reviewed by Magnus et al[70]. The final step will reconstruct and refine all-atom models from the best coarse grained conformations. While further improvements are needed in all of these stages, reliable RNA tertiary structure prediction is slowly getting within reach.

7 Acknowledgments

This work was funded, in part, by the Austrian FWF, project “SFB F43 RNA regulation of the transcriptome”. We would like to thank Roman Ochsenreiter for proof-reading the manuscript.
Figure 4: A proposed RNA 3D structure prediction pipeline, as described in Section 6. In addition to the structure prediction (middle), additional experimental (left) and computational steps (right) can improve the prediction accuracy.
References


